The Thesis Committee for Christos George Bampis Certifies that this is the approved version of the following thesis:

Subjective Quality of Experience in Video Streaming

APPROVED BY

SUPERVISING COMMITTEE:

Alan C. Bovik, Supervisor

Haris Vikalo

Subjective Quality of Experience in Video Streaming

by

Christos George Bampis, B. E.

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN ENGINEERING

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2016

Dedicated to my family and my best friend John.

Acknowledgments

I wish to thank those that have helped me in completing this thesis. I would like to thank my advisor Alan C. Bovik for inspiring and guiding me throughout my studies. He has become a role model for me as a researcher and an academic. Also, I am very grateful to the Netflix Video Engineering team (Zhi Li, Anush K. Moorthy, Ioannis Katsavounidis and Anne Aaron) for supporting this work and introducing me into the real world challenges of video research. I am grateful to my undergraduate advisor and mentor at NTUA Petros Maragos. He was the first one to introduce me to the areas of Image Processing and Computer Vision and excited my research interests since then. I would also like to thank my labmates and friends at LIVE: Todd, Zeina, Janice, Deepti, Leo, Lark and Praful whose advice and encouragement was plentiful. Further, I would like to thank prof. Haris Vikalo for reading this thesis and providing valuable comments.

I would like to thank my good friends for all those valuable moments in life and especially John: he is a true brother to me.

I would like to thank my whole family and especially my parents George and Penelope and my sister Valia for being there for me always. I feel indebted to my grandfather Christos and my grandmother Stavroula for their love during my childhood years. I will always remember them. Last, I want to thank my biggest passion: poetry. It has always been my escape route.

Subjective Quality of Experience in Video Streaming

Christos George Bampis, M.S.E.
The University of Texas at Austin, 2016

Supervisor: Alan C. Bovik

0.1 Abstract

This work studies subjective video quality of experience (QoE) in video streaming applications[1]. Streaming content providers such as YouTube are increasingly deploying HTTP adaptive streaming (HAS) strategies, where the video content is first divided into data chunks then encoded at different bitrates. Based on the estimated network conditions, a client can determine which bitrate will be used for the segment to be played next.

By studying previous works on subjective video quality, we first demonstrate that most subjective studies and QoE datasets are not driven by practical network constraints and may not be appropriate for real-world video streaming applications. Next, we describe our research efforts towards bridging this gap by designing the LIVE-Netflix QoE dataset, which simulates realistic network conditions in a typical video streaming scenario, using long video sequences.

[1]Parts of this work will be submitted to IEEE Transactions on Circuits and Systems for Video Technology.

Table of Contents

Chapter 1

Introduction

1.1 Introduction

Global mobile data traffic grew 74% and mobile video traffic accounted for 55 percent of total mobile data traffic in 2015 [1]. According to the Cisco Visual Networking Index and global mobile data traffic forecast, mobile data traffic will grow 8-fold from 2015 to 2020, which constitutes a compound annual growth rate of 53%. Given this large and growing volume of mobile video data, video streaming providers such as Netflix, Youtube and Hulu are processing, storing and delivering vast amounts of video data on a daily basis. Given the exploding use of mobile video devices and the tremendous network bandwidth demands of streaming users, the biggest challenges in video content delivery are to create better network-aware strategies to improve end-users quality of experience (QoE). In this direction, HTTP Adaptive Streaming (HAS) is being used by content providers as a way of dealing with network fluctuations.

1.2 HTTP Adaptive Streaming

The main idea behind adaptive video streaming is that the high bitrate video content is encoded at multiple bitrates and fragmented to small HTTP-based file segments of 2 to 10 seconds. A manifest file is used to inform the streaming client about the available bitrates and the segments of the streams.

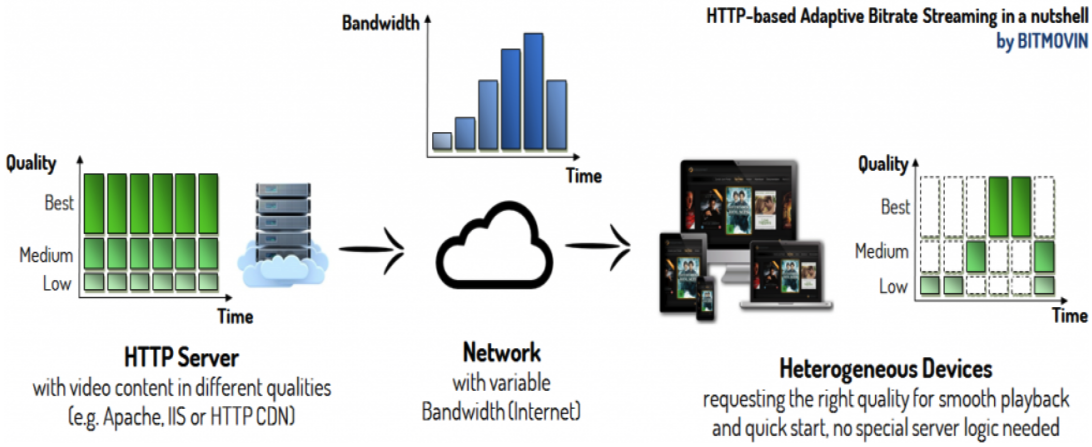

Figure 1.1: Overview of MPEG-DASH from [2].

Depending on the available network resources, the client side requests the highest available bitrate.

MPEG-DASH (Dynamic Adaptive Streaming over HTTP) is a widely known international standard for adaptive bit-rate HTTP-based streaming. An overview of this technique is shown in Fig. 1.1. The key benefits of MPEG-DASH are that it can reduce the number of startup delays and rebuffering events during the video and that it adapts to the bandwidth conditions of the client in a continuous way [2]. MPEG-DASH is also device independent since the client side can be anything: a home TV, a game console or a web browser.

Streaming video content providers like Youtube or Netflix are increasingly deploying such technologies to maximize the end user's QoE. Clearly, adaptive bitrate selection (ABR) algorithms have to balance between two competing factors [3]. On the one hand, they seek to maximize the video quality by selecting the highest possible bitrate. On the other hand, they try to minimize rebuffering events where the video playout stops completely and

a "loading" icon appears. However, if the ABR algorithm always picks the highest available buffer, it is likely that the playback buffer will be empty and rebuffering will occur. Therefore, an ABR algorithm has to consider a number of network-aware parameters: the previous network state, the buffer state and estimates of the network conditions within a small time window in the future (if available). Since the network bandwidth is temporally varying and hard to predict, this approach may lead to impairments including re-buffering events and/or compression artifacts.

1.3 Subjective Video Quality

Given that the end goal of every content provider is to maximize the end-user's QoE while mediating parameters to accommodate network changes and bandwidth throttling, perceptually-driven optimization strategies are the key to solving the resource allocation problem. However, QoE modeling is still far from being an easy task. The low-level human visual system (HVS) is complex and driven by non-linear processes not yet well understood. There are also cognitive factors that influence perceived QoE, adding further layers of complexity. For example, subjective QoE is affected by recency: more recent QoE experiences may have a higher impact on currently perceived QoE [4]. Subjective testing is an established way of analyzing subjective QoE under different scenarios and settings. In this work, we will study previous works on subjective QoE and discuss our own research efforts towards this direction.

1.4 Objective Video Quality

Designing subjective studies is time consuming and, in most applications, subjective data may not be available. To automatically predict video quality, many objective video quality metrics have been proposed. There are two broad categories of video quality metrics: full-reference (FR) and no-reference (NR) methods [5]. The former assumes that both the distorted and the pristine videos are available while the latter assumes that only the distorted video is given. In both cases, video quality is measured on videos of normal playback, i.e. on videos without any rebuffering. Clearly, video quality during normal playback is not the only QoE-relevant factor for streaming applications, where both rebuffering and video degradations may occur.

A different approach is to consider the video Quality of Service (QoS) by quantifying the effect of playback interruption on subjective QoE. Various approaches have studied the key parameters of rebuffering, such as the rebuffering number and duration and its location in the video (initial delay vs. rebuffering). Again, many of these approaches do not consider the combined effect of rebuffering and video quality degradation. In this work, we will discuss the tradeoffs between rebuffering and compression artifacts and motivate the need for deploying general QoE-aware models.

Chapter 2

Previous Works on Video Quality Assessment

2.1 Subjective Video Quality Assessment

Many subjective studies have been developed in order to better understand subjective video QoE. Seshadrinathan *et al.* [6] designed the LIVE Video Dataset using short video sequences (each 10 sec. long) afflicted by MPEG-2 and H.264 compression as well as network-related distortions such as transmission over error-prone IP and wireless networks. MPEG-2 and H.264 compressed bitstreams exhibit typical compression artifacts such as blocking which are fairly uniform across the entire video sequence, both spatially and temporally. By contrast, network losses lead to transient distortions which appear like "glitches". Figure 2.1 shows an example of an H.264 compressed frame and a frame transmitted over a (simulated) error-prone IP network. By simulating the aforementioned visual degradations, Seshadrinathan *et al.* gathered summary subjective ratings of visual quality: each subject gives a single number of his or her perceived video quality for each video. Using these subjective ratings, one can then evaluate and design new objective video quality prediction models such as MOVIE [7] and ST-MAD [8]. In the next section, objective video quality prediction models will be discussed further.

There have been many other video quality datasets that have been designed. For example, Moorthy et al. studied the effect of quality switches and rebuffering on video sequences displayed on a mobile device [9]. To gain

Figure 2.1: An example of video degradation. Left: H.264 compression; Right: IP loss.

a better understanding of how humans integrate their continuous time QoE experiences to a single summary QoE rating, they also gathered continuous subjective scores. It was suggested that "switching to an intermediate rate before switching to a higher rate is preferred over multiple large-magnitude rate switches". Also, they found evidence demonstrating the recency effect [4], by observing that the end quality of the video had a definite impact on the final perceived quality. Other subjective studies have also been designed to understand subjective QoE as in [10] where a streaming dataset was developed and in [11] where the interaction between rebuffering and compression was investigated. However, these studies do not reflect typical video streaming situations, where subjects view videos that could be minutes long. Hence, it is not possible to analyze long-term memory effects as they relate to critical factors affecting subjective QoE

Longer video sequences were considered in [12], where video delivery over HAS was simulated on tablet devices. The authors studied combinations of bitrate changes and re-buffering events, but their analysis was limited to 6 sequences, 3 playout scenarios and 26 subjects. Longer video sequences were also used in [13] using video contents ranging between 30 seconds and

a minute. The authors studied the effect of re-buffering events as functions of location and density in a video sequence. However, temporal ratings were not collected, hence their analysis was based only on a final summary rating. As we will show later, using the final rating only introduces recency biases. The study of temporal pooling techniques in [14] also included longer video sequences, and concluded that current temporal pooling strategies are mostly effective on short videos. On the long videos they used, simple mean pooling was superior to all other methods. However, they only used two video contents in their analysis. Finally, we note that in almost all previous subjective video quality studies, subject rejection strategies were based only on final scores or were conducted on a per frame basis. We argue that such methodologies are inappropriate when gathering temporal scores, particularly when studying the complex temporal effects that affect subjective QoE.

2.2 Objective Video Quality Assessment

Objective prediction models of subjective QoE play a very important role in perceptually-driven resource allocation problems. Full-reference (FR) methods rely on comparing the distorted video with the given pristine video. Frame-based image quality assessment (FR-IQA) methods can be naturally deployed for VQA, e.g. PSNR, PSNRhvs [15], SSIM [16], MS-SSIM [17] and GMSD [18]. However, these frame-based methods do not take into account the temporal aspects of video distortions. Therefore, temporal FR-VQA methods have been developed such as VQM_VFD [19], the powerful MOVIE index [7], ST-MAD [8], the learning-based VMAF [20] and FLOSIM [21]. An alternative to full-reference models, are reduced-reference models, such as STRRED[22]. STRRED is an information-theoretic approach to VQA that builds on the

innovations in [23], [24]. It achieves quality prediction efficiency without the need to compute motion vectors unlike [25], [7].

No-reference (NR) VQA has also been deeply studied [26]. Frame-based NR methods like NIQE [27] can be considered, but their predictive performance is usually very low. Many distortion-specific NR VQA methods [28], [29], [30] have been designed to predict the effect of domain-relevant distortions on perceived quality. In a general model [31], a natural scene statistics model in the DCT domain was used to train a support vector regressor to predict the effects of packet loss, MPEG-2 and H.264 compression. VIIDEO [32] generalizes further by relying only on statistical regularities of natural videos, rather than on subjective scores or prior information about the distortion types. However, the NR VQA problem remains far from an ultimate solution.

2.3 Objective Video Quality of Service

Besides video quality degradations due to compression, there can be other network-related distortions, such as transient video "glitching" due to transmission over error-prone wireless networks or delays (start-up delay or rebuffering). HAS uses TCP as its transfer protocol hence only rebuffering events and start-up delays due to throughput/buffer limitations are prominent. Under this context, the video Quality of Service (QoS) is causally related to QoE [33]; hence various works have focused on quantifying the effects of playback interruption on subjective QoE. While the effects of rebuffering on QoE are not yet well understood, various studies have shown that the duration, frequency and location of rebuffering events severely affects QoE [34], [13], [35], [36]. By making use of global rebuffering statistics, Quality of Service (QoS)

models such as FTW [37] and VsQM [38] have been proposed. Mok *et al.* expressed QoE as a function of frequency and length of rebuffering events [39]. More recent efforts [35] have sought to both model the effects of rebuffering on user QoE, and to integrate them with models of recency [4].

2.4 Cognitive Aspects of Subjective QoE

One of the reasons why subjective QoE is hard to analyze and predict, is the fact that video quality degradations and rebuffering events are not the only contributing factors. When making QoE evaluations, humans demonstrate a number of cognitive-driven characteristics, such as recency [4]: more recent experiences contribute more on the perceived QoE. By contrast, the primacy mechanism may also determine subjective QoE: humans tend to recall events that occurred at the beginning of a series of events [40]. Also, there can be other non-linearities in the way humans make QoE decisions especially when we consider time-varying QoE evaluations: humans have different response rates to the visual stimuli that vary over time. Throughout this work, we will be revisiting some of these cognitive QoE-related factors in greater detail and, by gathering subjective data and studying them, will attempt to better understand the contribution of all these factors.

2.5 The Need for Better Subjective Data and General QoE-Aware Models

Based on our analysis on the previous sections, it is clear that there are two main directions that are yet to be explored. First, most of the previously designed subjective studies suffer from at least one of the following:

1. a small number of contents, playout patterns or number of subjects
2. lack of practical network constraints for streaming applications
3. use of short video sequences that do not capture long term temporal effects
4. not including both temporal and final summary ratings
5. not deploying temporal subject rejection methods

Meanwhile, most QoE models have either been designed for videos suffering from compression artifacts or from rebuffering, but not both. This is partly due to the unavailability of suitable subjective data, along with the difficulty of combining objective video quality models and rebuffering-related information into single QoE scores. In [41], FR quality algorithms such as SSIM and MS-SSIM were combined with rebuffering information yielding the Streaming Quality Index (SQI). However, they assumed that the effect of each rebuffering event is independent and additive which is contradictory to the model suggested in [35]. In [42], the authors fed QP values and rebuffering related features into a Random Neural Network learning model to make QoE predictions. However, their method was evaluated on only 4 contents and on short video sequences of 16 seconds, did not consider longer term memory effects and did not deploy perceptually relevant VQA algorithms. This suggests the need for larger streaming-oriented subjective datasets and algorithms which collectively build on perceptually driven VQA methods, rebuffering models and other QoE-aware features. In this work, we will focus on designing a subjective video QoE dataset suitable for streaming applications that satisfies the aforementioned points.

Chapter 3

The LIVE-Netflix Video Quality Dataset

When designing resource allocation strategies, content providers seek to answer the question: given a fixed amount of network resources, which strategy delivers the highest possible QoE? We consider here the tradeoffs that occur on end users' QoE when mediating between re-buffering events and bitrate reduction under a mobile low bitrate regime. To do so, we designed a set of realistic playout patterns, assuming the same network resources and same buffer limitations on each.

3.1 Network Constraints: Available Bandwidth and Buffer Limitations

To simulate realistic network conditions, consider the exemplar available bandwidth distribution shown in Fig. 3.1. The available bandwidth reaches a maximum of 250 kbps, there is a temporary bandwidth drop to 100 kbps of duration $d = 22.2167$ seconds until the bandwidth recovers to the original maximum value. This simple example of a bandwidth drop can be used as a building block to simulate models of more complex network conditions. Using this available bandwidth model, we derived eight test patterns based on the premise that the playout rate of the client side cannot exceed that of the available bandwidth. The only exception to this rule is when the client uses some of the available buffer. Next, we discuss the buffer usage aspects of the

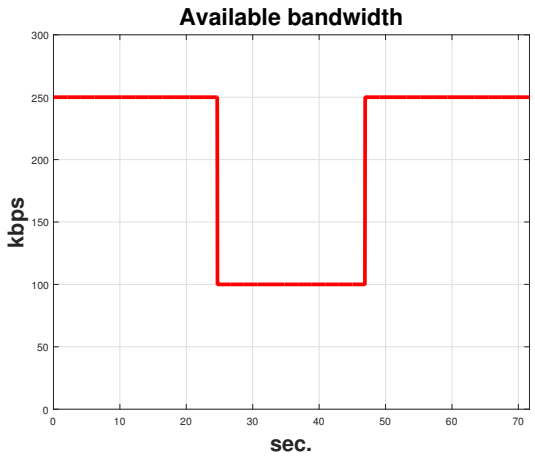

Figure 3.1: Available bandwidth model used in the LIVE-Netflix dataset. All of the test sequences were designed to consume the same amount of network resources (bandwidth).

designed patterns.

To ensure the practical worth of the derived sequences, it is necessary to take into account the available buffer size. As shown in [3], a buffer-based strategy can be a simple and useful way to reduce the number of re-buffering events and bitrate switches that occur. Clearly, there are three possibilities:

1. The playout rate is smaller than the available bandwidth; the buffer is being filled with more data.
2. The playout rate is larger than the available bandwidth; the buffer is being emptied.
3. The playout rate is equal to the available bandwidth; the buffer state does not change over time.

Given our network assumption, we also considered a specific initial buffer state for streaming, where the buffer of size B_0 was filled with video

chunks encoded at 250 kbps. We further assumed two possible initial buffer states: $B_0 = 1333$ kbits or $B_0 = 0$ kbits. The former scenario corresponds to "steady state" streaming where the initial buffer is filled, while the latter assumes that there is no initial buffer available. Further, we did not allow the buffer to be filled with video chunks encoded at different encoding bitrates: given a buffer filled with video chunks encoded at x kbps, no video chunk encoded at y kbps (where $x \neq y$) could fill the buffer until the entire buffer was first depleted. All patterns were designed so that the buffer is emptied at the end of the bandwidth drop shown in Fig. 3.1.

3.2 Playout Patterns

Based on this network scenario and possible values for B_0, we simulated the following client approaches (see also Fig. 3.2 for an overview):

0. A constant encoding bitrate of 500 kbps. This playout pattern assumes an impairment-free network condition where the bandwidth is rich enough to allow such a playout rate by the client. In this case, the buffer is not used at all. This pattern is the only one that does not satisfy the bandwidth and buffer constraints. It is used as a reference pattern.

1. One video chunk encoded at 250 kbps followed by an 8 sec. stall, followed by another 250 kbps chunk (see Fig. 3.3). The client drains the buffer completely before the re-buffering event occurs. Before the available bandwidth recovers, the client decides to resume playback after the 8 second rebuffer. By the end of the pattern, the buffer is emptied.

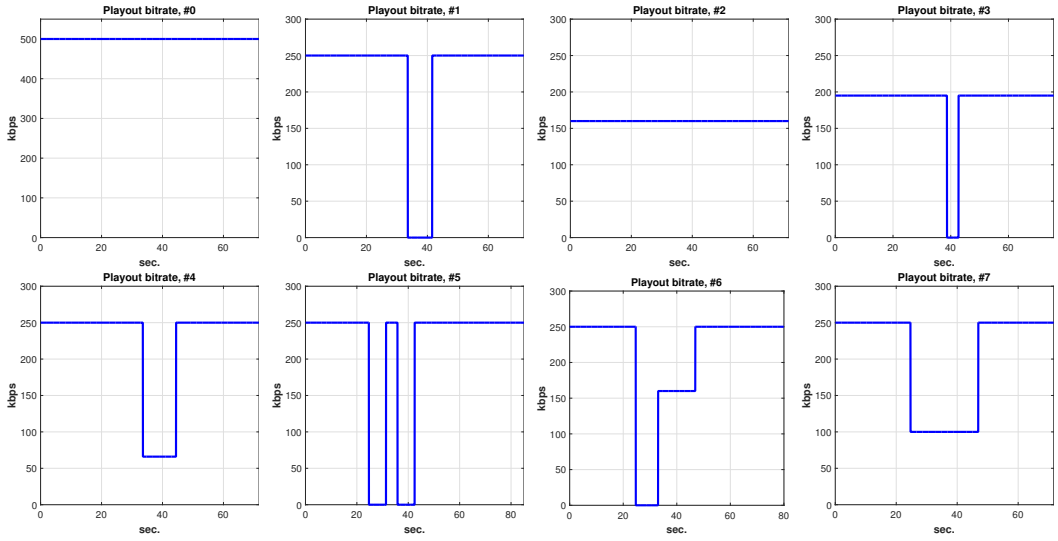

Figure 3.2: Playout patterns used in the subjective study. First row: patterns #0 until #3, second row: patterns #4 until #7. The horizontal axis corresponds to frame indices while the vertical corresponds to the playout bitrate in kbps.

2. A single video chunk of $R_2 = 160$ kbps. The client side is very conservative throughout the video playback by always picking a playout rate of R_2, so that there is no re-buffering and the available buffer is depleted.

3. One video chunk encoded at 195 kbps, followed by a 4sec. stall, followed by another 195 kbps chunk. Here, the client strategy is to reduce the re-buffering duration by half (4 sec.), by using a lower encoding bitrate. As before, during the re-buffering event, the client has a zero playout rate but an encoding bitrate of 100 kbps (equal to the available bandwidth) which allows the buffer level to partially recover and then be used to stream at 195 kbps before the bandwidth recovers (see also Fig. 3.3).

4. One video chunk encoded at 250 kbps followed by a 66 kbps chunk, followed by another 250 kbps chunk. This playout pattern is an alternative to pattern #1, where the client tries to avoid any re-buffering events by switching to a lower playout rate (66 kbps) than the available bandwidth (100 kbps) during the bandwidth drop.

By removing the assumption on the availability of the buffer on the client side ($B_0 = 0$), a second set of playout patterns can also be simulated. This set of patterns is likely to deliver lower QoE scores to subjects since more severe impairments have to be introduced to deal with the bandwidth drop.

5. One video chunk at 250 kbps, followed by a 6.66 sec. rebuffering event, followed by a chunk at 250 kbps, followed by another 6.66 sec. rebuffering event, followed by the last 250 kbps chunk. In pattern #5, the unavailability of the buffer leads to re-buffering. By filling some of the buffer, the client is able to play out for a small interval of time at 250 kbps until the buffer is depleted. This leads to the re-buffering event, which is followed by a recovery at 250 kbps playout over a small time interval until the bandwidth also recovers.

6. One video chunk at 250 kbps, followed by a 8.33 sec. rebuffering event, followed by a chunk at 160 kbps, then a final video chunk at 250 kbps. Here, the client seeks to avoid a second re-buffering event by a gradual bitrate recovery.

7. One video chunk at 195 kbps is followed by a chunk at 100 kbps and then another chunk encoded at 195 kbps. Here it is assumed that the client is immediately able to adjust to the network conditions by using a playout

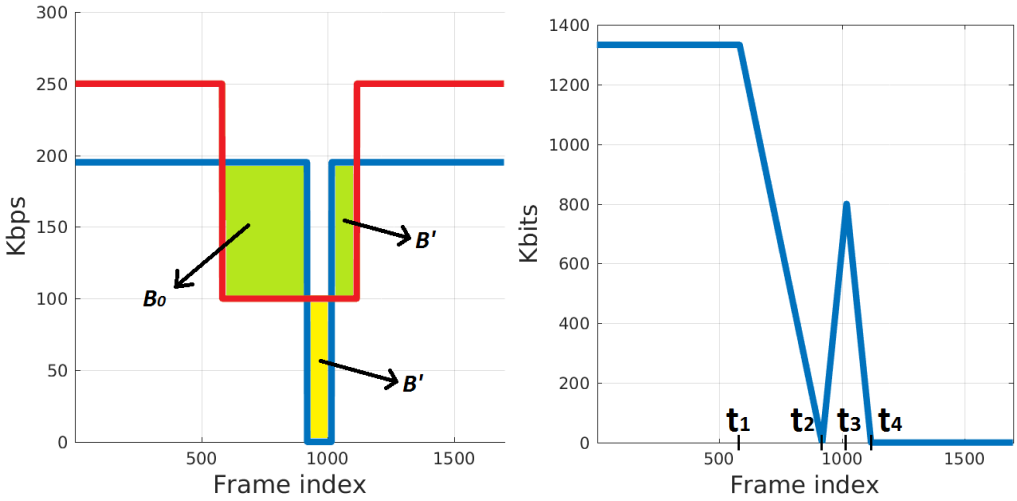

Figure 3.3: Left: Blue denotes the playout pattern #3 while red denotes the available bandwidth. The green areas correspond to buffer consumption while the yellow area indicates the buffer build-up. Right: Available buffer level over time for playout pattern #3, $[t_1\ t_2]$: buffer drainage, $[t_2\ t_3]$: buffer build-up, $[t_3\ t_4]$: buffer drainage.

rate that is always equal to the available bandwidth/encoding bitrate. This pattern may be the least practical among all the considered playout patterns. However, it is of interest to be able to study the subjective data resulting from such an "ideal" client reaction.

We now give an example of how the previous parameters were determined. We fixed the rebuffer duration for pattern #1 (see Fig. 3.2) to 8 sec. and the average bitrate for the client in pattern #2 to be $R_2 = 160$ kbps. Since there is no rebuffering event in pattern #2 but the available bandwidth is 100 kbps for d seconds, the client in #2 expends all of the available buffer B_0 in d seconds hence $(R_2 - 100)d = B_0$ yielding $B_0 = 1333$ kbits. Let t_b be the time interval after the available bandwidth drops until a rebuffering

event occurs in #1. Clearly, $t_b(250 - 100) = B_0$ since the client depletes all of the buffer before the playback interruption. During the rebuffering event, the buffer fills to $B_1 = 800$ kbits given 8 seconds of rebuffer at the available bandwidth of 100 kbps. The client chooses to start the playback t_a seconds before the available bandwidth recovers hence $t_a(250 - 100) = B_1$ (due to our assumption that all patterns will eventually deplete all of the available buffer). Therefore, $t_a = 5.3333$ sec. and $d = t_e + 8 + t_a \approx 22.2167$ seconds.

3.3 Encoding Pipeline

First, the high quality video stream (H.264 format file) is combined with the audio stream and placed in an mp4 container without encoding. Then, following the specific network-simulated pattern, the .mp4 file is divided into a number of different chunks each with a different encoding bitrate. For example, consider the general case of pattern #6, i.e., a pattern containing both bitrate changes and a re-buffering event.

We developed an encoding pipeline that generates the different parts of the final video and appropriately concatenates them based on an encoding map that indicates the time intervals of every quality level, the location and the duration of each re-buffering event as follows: enc $<$ start $>$ $<$ stop $>$ $<$ bitrate $>$ stall $<$ start $>$ $<$ duration $>$, where time was measured in seconds and the bitrate in kbps. The encoding resolution was based on the used bitrate, the encoding profile was set to high and the PAR was set to 4:3.

Using this encoding map the encoding process was carried out as follows (see Fig. 3.4). First, the source video and audio streams were transferred from the Amazon Cloud and stored locally for further encoding. Next, the source video stream (in H.264 format) was decoded, yielding an uncompressed raw

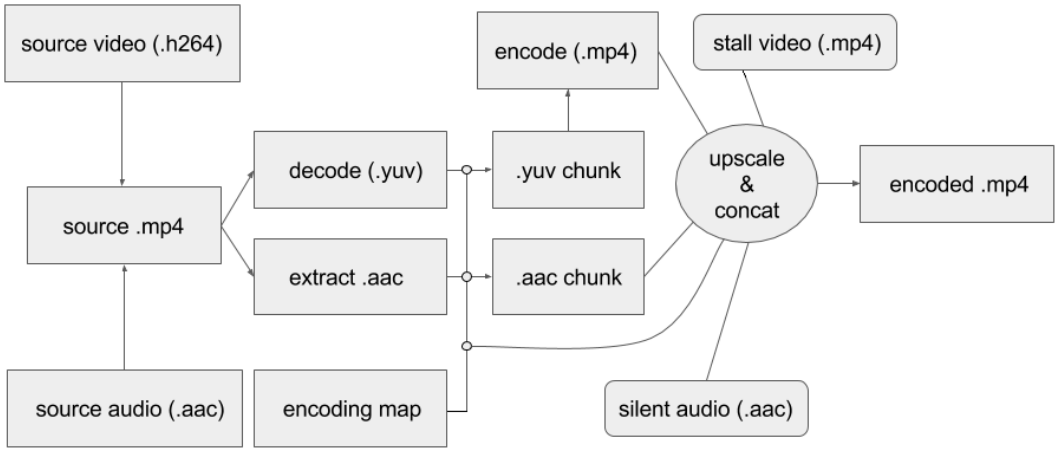

Figure 3.4: Encoding pipeline used to create the playout patterns.

.yuv file. The encoding map was then used to split the .yuv file in a frame-accurate manner, yielding .yuv chunks, e.g. three chunks for pattern #6. A two pass encoding step using ffmpeg was then used to encode the .yuv files into .mp4 format. For pattern #6, this corresponds to two chunks encoded at 250 kbps and one encoded at 160 kbps. Meanwhile, the final frame of a video chunk immediately before a re-buffering event was used to generate a re-buffering video chunk. A customized "loading" icon was overlaid on this frame and appropriately animated to simulate the desired video re-buffering effect.

3.4 Source Contents

A set of 14 video test contents were used containing a wide variety of spatiotemporal characteristics. Of the 14 contents, 11 are Netflix videos including action scenes, drama, adventure, anime and cartoons. The remaining 3 contents were obtained from the publicly available Consumer Digital Video

Figure 3.5: Some frames from the LIVE-Netflix dataset. From left to right: ElFuente and Chimera sequences from the dataset.

Library (CDVL). A few frames from the video sequences are shown in Fig. 3.5. The test contents have a variety of frame rates and resolutions. For example, the ElFuente sequence has 4K resolution (4096x2160) and a frame rate of 60 fps, whereas most of the Netflix contents have 1080p (1920x1080) resolution and frame rates of either 24, 25 or 30 fps. To deal with this difference, the ElFuente sequence was downscaled to 1080p and the frame rate was converted from 60 fps to 30 fps.

Measurements of spatial and temporal complexity give a rough idea of the content variety in a subjective database [43]. Let F_n denote the luminance channel of a video frame at time n and $(i,\ j)$ the spatial coordinates of this frame. Next, consider the following simple Spatial Information (SI) and Temporal Information (TI) metrics [44]:

$$\mathrm{SI} = \max_n \left\{ \mathrm{std}_{i,j}[\mathrm{Sobel}(F_n)] \right\} \tag{3.1}$$

$$\mathrm{TI} = \max_n \left\{ \mathrm{std}_{i,j}[M_n(i,j)] \right\} \tag{3.2}$$

where $M_n(i,j) = F_n(i,j) - F_{n-1}(i,j)$, $\mathrm{std}_{i,j}(.)$ denotes the standard deviation over all pixels (i,j), $\max_n$ denotes the maximum over all frames and Sobel(.) denotes the Sobel operator, which convolves the image with a 3×3 spatial filter and calculates the horizontal and vertical edges. As shown in Fig. 3.6, the video content we use widely spans the SI-TI space [44].

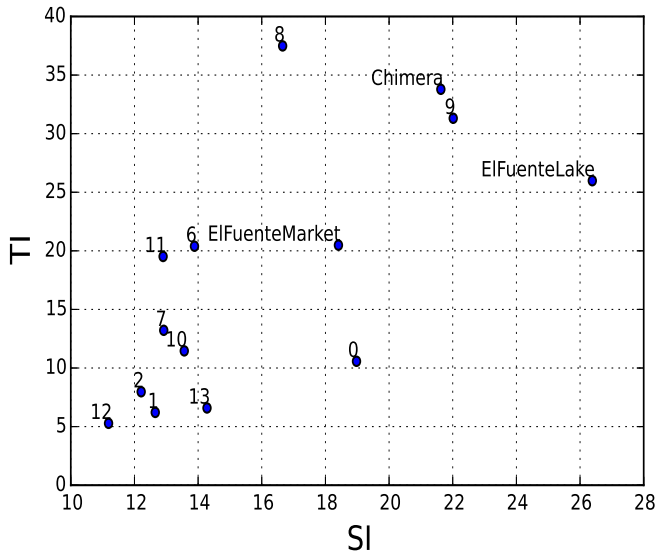

Figure 3.6: Spatial Information (SI) plotted against Temporal Information (TI) for the 14 video test contents in the LIVE-Netflix dataset.

3.5 Subjective Testing Design

A single-stimulus continuous quality evaluation study [45] was conducted over a period of three weeks at The University of Texas at Austin's LIVE subjective testing lab. Although we included reference pattern #0 among the viewed playout patterns, it did not serve as a "hidden reference." Since the generated video patterns are of different duration because of the introduction of re-buffering events, computing temporal Differential Mean Opinion Scores (DMOS) was not possible.

Due to necessary limitations on the duration of a subjective study, video QoE studies invariably must limit the number of different contents that are shown. When using longer video sequences, this is even more challenging.

Driven by a desire to deploy as diverse and large set of contents as possible, we employed the following strategy. Each subject was assigned 11 contents (of the 14) in a circular fashion e.g. if subject i as assigned contents 1 through 11, then subject $i+1$ watched contents 2 through 12. This could result in a slightly different number of temporal and final scores per content, but given the large number of subjects, we deemed this to be a statistically insignificant difference. All 8 playout patterns for these 11 contents were displayed to the subject only once. In order to remove any memory effects, we randomly shuffled the contents and the corresponding playout patterns while ensuring that the same content was not consecutively displayed to a subject in any session. Visual fatigue is an important consideration when designing subjective studies, so we split the study into three sessions, spaced by at least 24 hours to minimize subject fatigue [45]. Each session contained video content at most 35 minutes long, and the overall duration of each session was about 45 minutes. The first session was 5 minutes longer due to the training process.

Android Studio was used to modify an earlier version of the human subject interface used in [9], which was made available to us by the authors. Using the previously described encoding pipeline, the generated .mp4 files were displayed on a Samsung S5 mobile device with a 1080p resolution and 5.1" screen size. This device had no problems playing the videos which were stored locally on an external SD card. The use of an external SD card did not introduce any latency when displaying the videos. The mobile device was not calibrated, but the brightness level was held constant at approximately 75% of maximum throughout the study. In order to supply a more realistic viewing experience, we decided to supply the associated audio without any introduced distortions (other than rebuffering events aligned with the videos

and compression artifacts at bitrates as low as 66 kbps) on the device's speakers when playing each video content.

Each subject participated in three sessions. The first session consisted of a training process, where the subjects signed a participation consent form and read a set of instructions guiding them through the study process. These instructions were also clarified verbally. No formal visual acuity test was performed, but the subjects verbally verified that they had normal or corrected-to-normal acuity. If a subject normally used corrective lenses when watching videos, they were asked to use them during the study. The subjects were asked to rate both their continuous and their overall QoE based on everything that they viewed on the screen. They were also asked not to make QoE judgments based on the level of interestingness of the video content or the audio quality. To remove any rating biases, the subjects were informed that there were no right or wrong answers in the experiment.

The subjective testing procedure during the first (training) session is described next. First, the subjects were introduced to the interface and the different video impairments they would be exposed to. Three different video contents, each with a different playout pattern were displayed as each subject became familiar with the testing interface. These contents were the same for all subjects but were not among the test contents used to gather the subjective data. An example of this interface is shown in Fig. 3.7. The subjects used an external mouse to place their QoE ratings and to navigate through the interface. After the first session, no training videos were shown, since subjects were assumed to be adequately familiar with the testing procedure and interface.

The video sequences in each session were displayed one after the other and a continuous scale rating bar was displayed at the bottom of the mobile

device screen. The ratings on the continuous (Likert) scale ranged from 0 (Bad) to 5 (Excellent). After each video finished, the subjects were asked to give an overall rating of their QoE using the same rating bar. Then, a screen prompt allowed the subjects to take a short break before they could initiate the playout of the next video. Examples of these steps can be seen in Fig. 3.7. The sampling rate on the continuous scores was such that one score was measured per frame. Given the different frame rates of the input sequences, we parameterized the number of samples per video content depending on each video's frame rate.

Figure 3.7: Subjective testing interfaces. Left: continuous QoE scoring; Right: final scoring.

3.6 Subjective Data Processing

We collected subjective data from 56 subjects and a total of 4928 continuous scores together with the corresponding final scores. Then, z-score normalization was applied on a per session and per subject basis to account for differences in the use of the rating scale by each subject, for each of the 3 viewing sessions. Let $s_{ijk}(t)$ and f_{ijk} denote the continuous scores and the final score assigned by subject i to video j during session k and let t denote

the frame number. Note that the set of all j videos viewed by subject i may not have been exactly the same for another subject i'. Consider the following operations:

$$\hat{s}_{ijk}(t) = \frac{s_{ijk}(t) - \mu_{s,ik}}{\sigma_{s,ik}} \tag{3.3}$$

$$\hat{f}_{ijk}(t) = \frac{f_{ijk} - \mu_{f,ik}}{\sigma_{f,ik}} \tag{3.4}$$

where $\mu_{s,ik}$, $\mu_{f,ik}$ are the mean continuous and final scores assigned to all videos at session k of subject i and $\sigma_{s,ik}$, $\sigma_{f,ik}$ are the corresponding standard deviations. No DMOS computation was applied.

Using the subjective data in the form of z-scores, the next step was to apply subject rejection strategies to identify potential outliers in the rating process. In video quality studies with longer videos, it is possible that subjects demonstrate less motivation and/or attention on some videos than on others. We believe that subject rejection methodologies based only on final scores are questionable for the following two reasons. First, if some subject is rejected based on only a single score per video but then is also discarded from all other video sequences he or she viewed (as is typically done), such a strict rejection criterion may needlessly reduce the amount of data. In our case, applying subject rejection only on the final scores as suggested in [45], [6] led to 7 subjects being marked as outliers. Since we focused on the temporal effects of subjective QoE, we considered it sensible to enrich the subject rejection strategy by taking into account the temporal dimension of subjective QoE.

In our preliminary design of temporal subject rejection schemes, we experimented with simple heuristics. First, we applied the frame-to-frame equivalent of final score rejection [45], [6], [46] which yielded inconsistent results. We believe this was due to the fact that introducing both dynamic bitrate

changes and re-buffering events led to more complex subject reactions with different response and lag times. An alternative approach is to apply a simple thresholding method: discard subjects that are un-responsive during any re-buffering event. However, we encountered instances where subjects did not react to a re-buffering event but were very unforgiving of a second re-buffering. This observation led us to avoid using such simple *ad hoc* methods.

We instead deployed a more sophisticated dynamic time warping (DTW) [47] strategy on the subjective ratings to identify similarities in aligned *temporal* subject responses. Subjects that were completely un-responsive during a time period where most of the other subjects reacted were noted. Consider subject i and the temporal rating waveform s_{ij}, where j denotes a video content using one of the 8 playout patterns. We collected all warped distances between subjects i and k, i.e., $d_{ik} = \text{DTW}(s_{ij}, s_{kj})$, where d_{ik} denotes the temporal misalignment between subjects i and k. This is a measure of dissimilarity: a large d_{ik} could mean that subject i reacted very rapidly to some stimuli whereas subject k reacted more slowly. Subject ratings having large distances from most of the others can be thought of as unreliable. As we have already explained, however, only per video rejection decisions were made, i.e., if subject i had unreliable ratings on some video j it did not imply rejection of all the other subject's ratings. Note that before computing the DTW distances, every waveform was scaled to the range $[0,\ 1]$.

Computing the d_{ik} yielded a matrix $\mathbf{D} = [d_{ik}]$ describing the temporal misalignments between all subjects that viewed video j. Since the DTW distance is symmetric, we computed only the upper triangular part of the matrix and set $d_{ii} = 0\ \forall i$. Then, the sum of the DTW distances across the rows (or columns) of $\mathbf{D}$ may be considered to be a measure of how unreliable a sub-

ject is: a large accumulated distance implies a subject whose responses were consistently mis-aligned with respect to other subjects on the video.

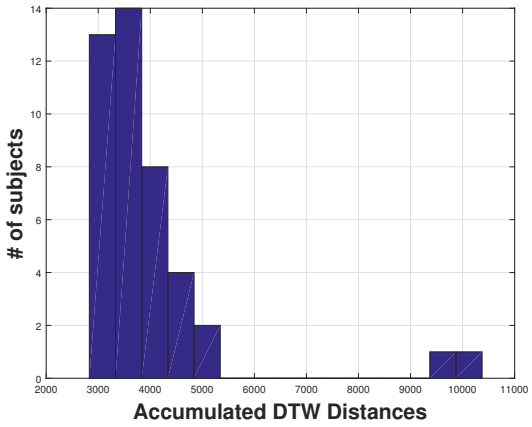

Figure 3.8: Distribution of accumulated DTW distances computed on one test video. The rightmost subjects have a higher chance of being outliers.

In Fig. 3.8, the distribution of accumulated DTW distances is shown for one of the test videos. The horizontal axis corresponds to the sum of the rows in $\mathbf{D}$, while the vertical axis indicates the number of subjects having the corresponding DTW distance. The distribution of accumulated distances is skewed to the right, making outlier identification more challenging. A standard technique is to apply Tukey's boxplot [48] rule, viz., mark all observations that are smaller than or that exceed 1.5IQR as outliers, where IQR is the interquartile range Q_3-Q_1 where Q_1 is the 25th percentile and Q_3 the 75th percentile. However, this rule assumes an underlying normal distribution. To address the skewness of the data distribution, we can either transform the data using an appropriate transformation (e.g. a Box-Cox [49] transformation) or use an adjusted boxplot technique like the one in [50]. We used the adjusted boxplot method. Then, an observation is considered to be an outlier if it lies

outside the interval:

$$[Q_1 - h_l(\text{MC})\text{IQR}\ Q_3 - h_u(\text{MC})\text{IQR}] \tag{3.5}$$

where h_l and h_u are functions of the medcouple (MC), which is a skewness measure [50]. We used the exponential model proposed in [50] i.e. $h_l = 1.5 \exp^{\alpha \text{MC}}$ and $h_u = 1.5 \exp^{\beta \text{MC}}$, where α and β are weighting factors. We picked $\alpha = -4$ (default value) and $\beta = -1$ since the DTW distributions are right skewed, and a small value of β produced a more robust estimator. Using this skewness-driven boxplot, we identified potential outliers on each test video and removed them from the collected data.

3.7 Analysis of Summary Scores

We next discuss how we analyzed the subject scores using summary scores. First, we considered the overall distribution of the final MOS before z-scoring. Figure 3.9 shows the distribution of raw final MOS. It can be observed that the scores varied over the interval [1.5, 4.5], hence the entire scale [0, 5] was not used. However, the subjects were not prompted to use the entire scale, since this could introduce bias. Instead they were allowed to give their natural responses. Also, note that patterns #1 to #7 were given similar MOS scores.

In typical streaming applications, subjects are exposed to long video sequences, and events that occur early on may have less effect on the overall rating given by a subject. This is known as the "recency effect" [4] where recent events more heavily influence the current perception of one's viewing experiences.

To examine these biases further, we conducted a preliminary statistical analysis to determine whether the playout patterns were actually (final) scored

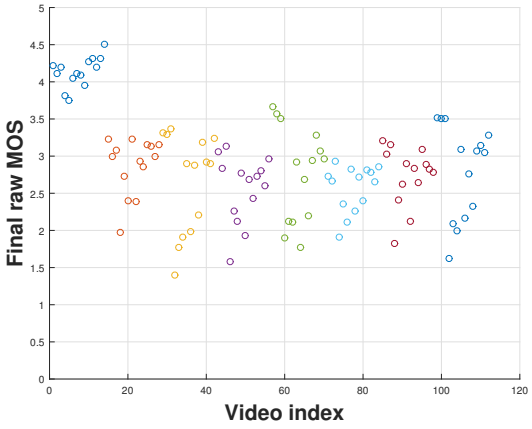

Figure 3.9: Raw MOS for all 8 patterns. Only pattern #0 is significantly different from the other 7.

differently by the subjects. We verified that the score distributions were not very skewed, then applied the Wilcoxon ranksum test (using a significance level $\alpha = 0.05$). We observed that, in many cases, the statistical comparisons between the final scores assigned to the playout patterns yielded statistically insignificant differences. This could be explained by recency (latest experiences matter for retrospective evaluations) and the duration neglect effect [4]: subjects may lower their temporal scores if a long lasting video impairment occurs. However, even if they did recall the duration of an impairment, they tended to be insensitive to its duration when making retrospective (overall) QoE evaluations. Also, note that by the time the subjects were asked to give an overall evaluation of each test video, more than 15 or 20 seconds of the 250 kbps playout had occurred. Given the tendency of subjects to evaluate videos based on more recent experiences, the test videos were possibly rated in response to the most recent video behavior.

If one is seeking a simple and direct QoE analysis, then it would seem desirable to obtain a single QoE value for each test video. Since the final

scores are affected by recency and duration neglect, we used simple frame averaging on the temporal scores to obtain a summary rating of each test video. Unfortunately, averaging continuous subjective scores without first applying temporal alignment does not account for the temporal QoE behavior of each subject (such as subject response delays). However, the DTW is appropriate only for pairwise time-series alignment, and may not produce an output having the same duration as the original waveforms. In our search for a recency-insensitive summary rating, we found that simple averaging correlated well with the final scores, as seen in Fig. 3.10. This observation aligns with two previous subjective studies: one where the test videos lasted only 10 seconds [6] and one with longer videos [14].

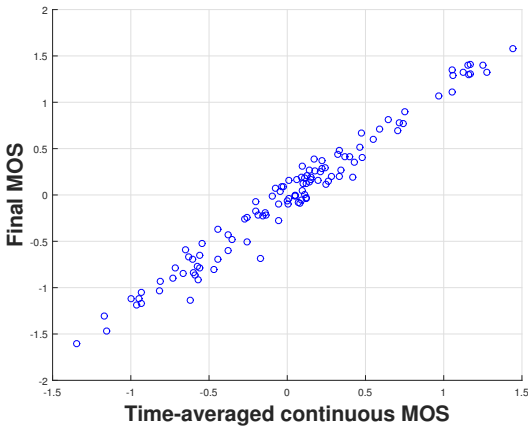

Figure 3.10: Scatter plot of the frame-averaged continuous scores (horizontal axis) against the final MOS (vertical axis) for all test videos.

Using the averaged scores as the summary ratings, we compared the playout patterns of each content as shown in Fig. 3.11. Clearly, patterns #5 and #6 were statistically inferior to the patterns from the first category (#1 to #4), since the available buffer was zero and fewer bits were spent; hence there was more rebuffering and/or lower bitrate values. By comparing #4 with

#1 and #2, two observations may be made. First, transient bitrate drops (#4) were not always preferred over rebuffering. In fact, we observed that the outcome of the statistical comparison depended on the level of compressibility of each content. For less compressible contents that require more bits (e.g. due to high motion), rebuffering was preferable, while for less compressible ones rebuffering was undesirable. However, a consistently low bitrate value (to avoid rebuffering), as in the "conservative" client strategy #2, was not tolerated by subjects. Further, subjects preferred a long rebuffering (#1) if it meant better quality elsewhere rather than the combination of a short rebuffering event combined with an intermediate recovery bitrate (#3).

Notably, pattern #7 had the best performance among patterns in the second category ($B_0 = 0$) and was comparable to #2 and #3. Again, this shows that subjects preferred transient bitrate drops. Surprisingly, #7 used fewer bits than #2 and #3 but yielded similar QoE. While #7 assumed an ideal client that could immediately adapt to the network conditions, this comparison demonstrates the merits of QoE-aware network policies: using fewer bits does not always mean that perceived quality is lower. However, we also observed that patterns #5 and #6 were statistically indistinguishable over all contents. This brings up another aspect of the subjective test's design: apart from recency, allocating the same number of bits under these circumstances could signify a similar retrospective QoE or summary rating. This underlines the need to exploit the temporal aspects of QoE, since the summary ratings reveal only some aspects of subject QoE.

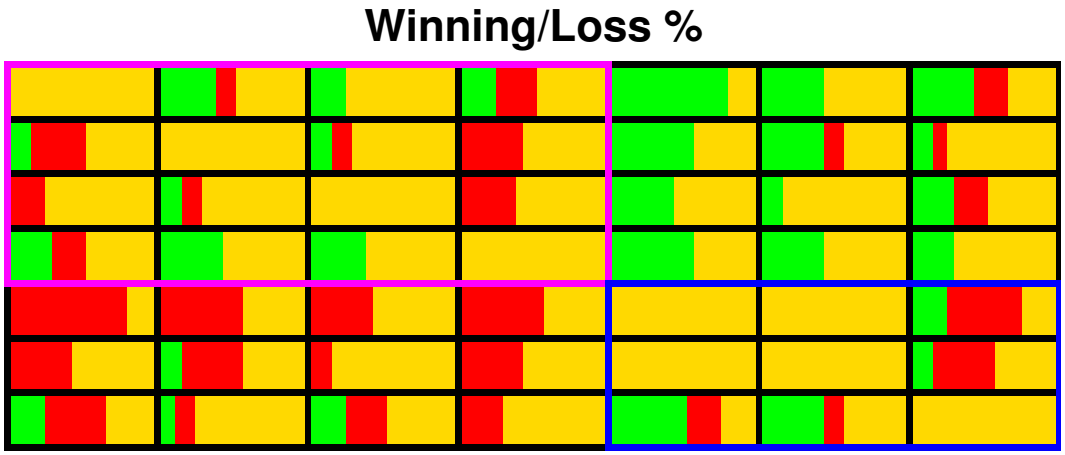

Figure 3.11: Statistical analysis of the averaged temporal scores for all patterns, represented as a 7×7 matrix. Each entry shows the winning percentage of the row compared to the column for all 14 video contents. Green shows the number of contents that the pattern in the row is QoE superior to the row, red shows the contents where the row is inferior to the column and orange shows that the row and column are indistinguishable. The purple box shows the comparisons only between patterns #1 to #4 ($B_0 = 1333$ kbits) and the blue box shows the comparisons only between patterns #5 to #7 ($B_0 = 0$ kbits).

3.8 Analysis of Temporal Scores

Temporal scores are a rich source of subjective QoE. Similar to the frame averaging used before, we performed frame averaging on the continuous subjective scores and show the result for several patterns in Fig. 3.12. We now focus on a comparison between patterns #1 and #7. Clearly, rebuffering severely and sharply damages the subjective QoE for all contents (in #1). Further, the QoE recovers at a slower pace than it originally dropped, suggestive of the hysteresis phenomenon: there is a lag between subjective QoE scores and current video quality or playback status. We earlier observed that subjects were not forgiving of rebuffering events. By contrast, when the bitrate dropped from 250 to 100 kbps, the subjective QoE reactions varied depending on each content. On scenes having higher spatiotemporal complexities, compression artifacts may be more visible and affect the QoE heavily and sharply, while others may not be affected to the same extent. Similar observations may be made for all patterns that contain at least one rebuffering event (where the video freezes and the rebuffering icon appears), which are obvious and unpleasant to viewers, whereas bitrate drops have a different impact on subjective QoE depending on each scene's encoding complexity (compressibility).

Notably, the constant encoding bitrate employed in #2 had a temporally varying effect on the perceived QoE. Given the long duration of the video contents and the different video characteristics present in each content (such as scene changes), it is clear that the subjects' QoE also changed over time even when the encoding scheme was static. This observation strongly supports a "per chunk" encoding strategy [51], where each video content is first split into short video chunks and then, based on the video complexity during this

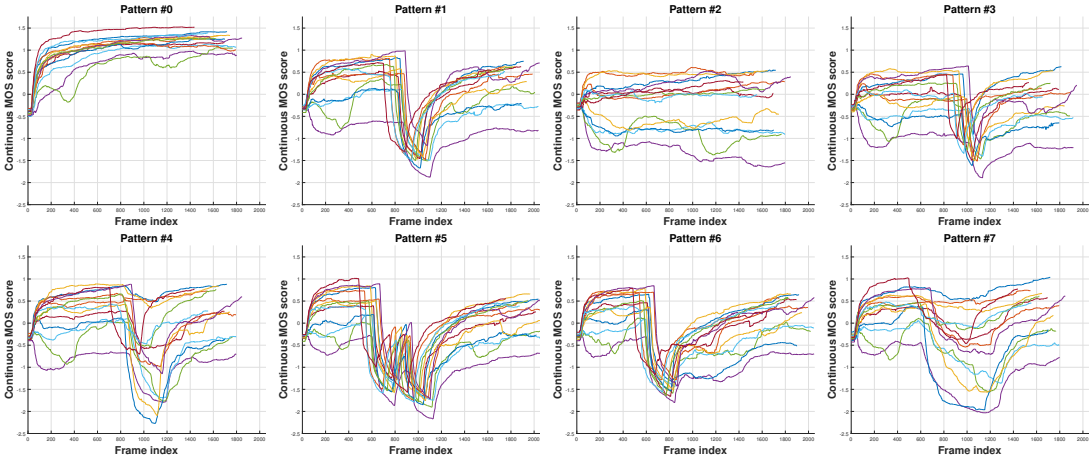

Figure 3.12: Temporal ratings across all contents for all playout patterns after subject rejection. First row: patterns 0 to 3; second row: patterns 4 to 7.

chunk, an appropriate encoding scheme can be chosen.

To investigate the interplay between rebuffering and compression artifacts in another light, we split the test contents into two sets based on their compressibility: Set 1 includes source contents which can be compressed easily and Set 2 those that are harder to compress. To determine the two sets we considered the following: contents with high motion and/or spatial complexity require more encoding bits, hence subjective scores would likely be lower on such sequences. To determine content complexity, the authors of [52] defined a criticality measure as the logarithm of the sum of the SI and TI indices.

However, here we are more interested in the degree of "compressibility" of the video contents. Given that the quality impairments of the otherwise very high quality videos being viewed are dominated by H.264 compression, an excellent measure of the "compressibility" of a video to a fixed bitrate are the scores of a high performance objective quality engine such as STRRED [22]. STRRED is an information-theoretic approach to VQA that builds on

the innovations in [23], [24]. It achieves quality prediction efficiency without the need to compute motion vectors unlike [25], [7].

To avoid any subjective biases due to content, we computed STRRED [22] between the original pristine video and #2 (constant encoding bitrate). The computed STRRED value (on the constant bitrate encodes) was a way of describing the compressibility (encoding complexity) of each video content: the higher the STRRED value, the less compressible the content was assumed to be. As we will show later, STRRED performed the best among the VQA models studied across the subset of video sequences without any rebuffering, hence it was deemed suitable for this purpose. Finally, as shown in Fig. 3.13, there are 5 contents (shown in red color) that have a relatively higher encoding complexity than the rest. Therefore, we considered those 5 contents as Set 2 while the rest were assigned to Set 1.

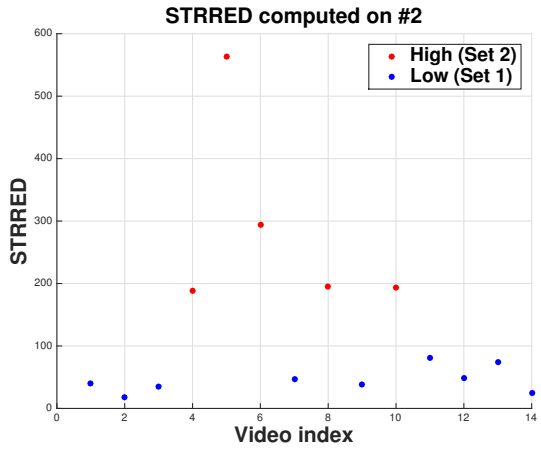

Figure 3.13: STRRED values between pattern #2 and the original source video for all 14 contents. Blue denotes more compressible contents while red indicates less compressible ones.

Next, we found the average (per frame) MOS score over all contents for

each of the 8 different patterns, as shown in Fig. 3.14. The effects of content complexity was evident: after a rebuffering event occurred, the QoE recovered more slowly for the contents in Set 2 (see green arrow in #6). Meanwhile, the videos in Set 2 tended to have larger standard errors against the videos in Set 1, since the increased encoding complexity may have led to a larger variance in the subjective QoE reactions. Overall, during normal playback, the contents in Set 2 have a lower QoE than the contents in Set 1.

We also observed the following interaction: a relatively long rebuffer event (as in playout patterns #1 and #6) led to larger drops in the reported subjective QoE on Set 1, as compared to Set 2 (see the black arrows in the plots for playout patterns #1 and #6). It is likely that the subjects were more annoyed by rebuffering events when they occurred during the playback of higher quality video content. A similar observation was also made in [11] using retrospective QoE ratings on short video sequences. However, for shorter rebuffering events (playout patterns #3 and #5) quality drops due to rebuffering between the two sets was similar. Notably, the second rebuffering in pattern #5 led to the opposite effect: given that one rebuffering event had already occurred, the quality drop on Set 2 was larger than the one for Set 1. This may be attributed to the effects of memory of a recent rebuffering event on currently perceived QoE.

By comparing patterns #1, #3 and #5, it is also evident that when the number or the durations of the rebuffering events increases, there is a larger drop in the temporal QoE scores. Again, these effects of rebuffering on the subjective QoE were harder to capture when we used the final summary ratings.

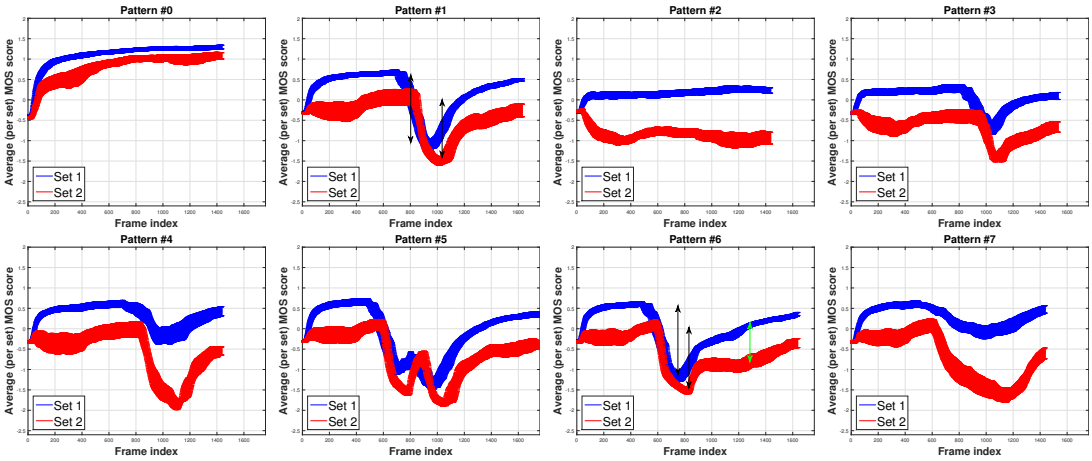

Figure 3.14: Averaged temporal ratings and standard errors for content Sets 1 and 2 for all playout patterns after subject rejection. First row: patterns 0 to 3; second row: patterns 4 to 7. Blue: Set 1; red: Set 2. Due to the different video lengths, we trimmed the axis of the plot to the duration of the shortest video sequence. The black arrows show the effect of rebuffering for the high vs. low complexity sets. The green arrow shows the different rates of QoE recovery for these sets.

3.9 Recency Biases and Non-Linearities

As already discussed, subject QoE might depend heavily on more recent experiences. To further investigate this claim, we performed local averaging on the temporal scores using a sliding window, then measured the correlations of those averages against the final scores. Let κ denote the size of the sliding window in seconds, τ be the total duration of a video and $\mu(a,\ b)$ be the average of the temporal scores from frame a to frame b and f be the final score assigned to that video. Figure 3.15 shows the SROCC between $\mu(a,\ b)$ and f using $\kappa = 10$ seconds. It is clear that local temporal averaging produced

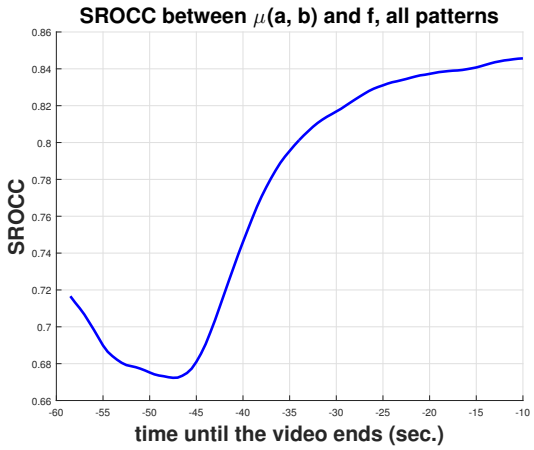

Figure 3.15: SROCC between the averaged temporal scores (over a 10 sec. window) and the final MOS.

stronger correlations over the more recent time intervals. This agrees strongly with the recency effect observed on the subjects' QoE.

Non-linearities in human responses to video quality are usually not considered in depth. Here, we are able to examine these effects given the richness of the collected temporal data. Fig. 3.15 shows that as the observation window is increased further into the past, the rank correlation decreases until

approximately 45 seconds, at which point it increases. This could be due to the fact that after the first 15 seconds most of the video impairments begin to occur, hence a local temporal window of "high disagreement" between subjects occurs as the impairments take place. By high disagreement, we refer to different response times between subjects, different recovery times and different use of the rating scale. Note that even after the z-scoring normalization, the subject ratings are still dependent on the rating behavior over time. We refer to both bitrate changes and re-buffering events during those time intervals as "events" where non-linearities in the human responses are activated and intensified. As a result, linearly combining the scores still produces non-linear measurements that do not correlate as well as when such events are not taking place.

3.10 Is Objective VQA Enough?

Most VQA algorithms are not applicable to frame freezes; hence video sequences with playback interruptions are usually not considered in objective quality analysis studies [9]. As a way of understanding how well these "standard" VQA models predict subjective QoE, we ask the question: "How well do VQA algorithms perform on video sequences with playback if applied only on the normal playback frames?". To answer this question, we considered the set S_q of videos without any rebuffering, the set S_r of videos having at least one rebuffering event and the whole dataset ($S_{all} = S_q \bigcup S_r$). Clearly, S_r and S_q are disjoint. Then, we applied various quality metrics on S_q and S_{all}. We compared several leading full reference (FR) and no reference (NR) image (IQA) or video (VQA) quality assessment algorithms [26], [5]: PSNR, PSNRhvs [15], SSIM [16], MS-SSIM [17], NIQE[27], VMAF [20], STRRED [22] and GMSD

[18]. When applying them on the videos in S_q, we calculated the quality scores only on normal playback frames and measured the correlation with the final scores after subject rejection. For PSNRhvs we used the publicly available Daala [53] implementation and for the other methods we used the available implementations. All models were applied on the luminance channel of the test videos. The results are tabulated in Table 3.1.

Table 3.1: Spearman's rank correlation coefficient (SROCC) for various image/video quality assessment algorithms (IQA/VQA) after performing mean pooling on the no rebuffering subset (S_q) and on the whole dataset (S_{all}).

IQA/VQA metric	S_q	S_{all}
PSNR (IQA, FR)	0.5561	0.5152
PSNRhvs [15] (IQA, FR)	0.5841	0.5385
SSIM [16] (IQA, FR)	0.7852	0.7015
MS-SSIM [17] (IQA, FR)	0.7532	0.6800
NIQE [27] (IQA, NR)	0.3960	0.1697
VMAF [20] (VQA, FR)	0.7533	0.6097
STRRED [22] (VQA, RR)	0.7996	0.6594
GMSD [18] (IQA, FR)	0.6476	0.5812

As shown in the first column, NIQE unsurprisingly performed the worst since it is a frame-based NR model, while PSNR and PSNRhvs performed the worst across all FR algorithms, followed by GMSD. The results on S_{all} were much lower than on S_q; indicating that the tested IQA/VQA systems were unable to predict QoE as well when rebuffering events were present. Note that SSIM performed better than MS-SSIM and close to the best predictor (STRRED) on S_q. This suggests that the subjects were internally responding strongly to rebuffering events rather than evaluating quality only. Further, this shows that in the presence of rebuffering, objective video quality models become less reliable predictors of subjective QoE. This may also explain

why VMAF delivered relatively low performance: it was trained on different subjective data, where the subjects were exposed to video suffering only from quality impairments. This implies the need to develop more general QoE-aware methods.

Chapter 4

Future Work

We introduced network constraints and used the buffer size as equalization factors and considered eight different streaming approaches that clients can apply. We designed each of these approaches such that re-buffering events and dynamic bitrate changes were both considered; either in isolation or combined. We then carried out a subjective study where more than 50 subjects viewed and rated those playout patterns on a diverse set of Netflix and publicly available video content. We collected both continuous and final ratings from the subjects on a mobile device.

When using long video sequences, the overall (final) score is inadequate to develop subject rejection schemes. We introduced a novel way of subject rejection via a dynamic time warping scheme applied on the continuous subject scores. This allowed us to apply temporal subject rejection on a per video basis. We observed that averaged temporal subject scores were less affected by memory biases but were also still correlated with the final scores. Then, we used the averaged scores to identify statistical differences between the playout patterns.

Through a statistical analysis, we examined the interactions between rebuffering and bitrate changes. We found that rebuffering is more obvious to subjects across contents while bitrate drops may not always be unpleasant. Transient bitrate drops were preferred for more compressible contents even

when the selected bitrate was low while a consistently low bitrate (to avoid re-buffering) yielded lower subject scores. We analyzed temporal effects on subject QoE and studied how subjects integrated their instantaneous experiences into a single final score. Our analysis supported the recency bias on subjective QoE, where subjects tend to bias their responses towards their most recently viewed experiences when asked to given an overall score. An obvious takeaway from analyzing the final scores was that they are less informative of a subject's QoE when longer video sequences are studied.

Apart from the inherent memory effects that were studied, the data analysis also resulted in our achieving a better understanding of the non-linearities of human responses. In particular, we observed that subjects reacted in different ways i.e. they had different response times with respect to those impairments. This disagreement was observed as a temporary drop in the correlation between locally averaged continuous scores and the final score. We then focused on objective QoE prediction models by studying the performance of state of the art video quality methods. Clearly, their performance was poor which demonstrates the need for deploying more general QoE-aware models.

We described a human study that focused on the temporal aspects of subjective video QoE under various network, buffer and low bitrate constraints. However, the temporal aspects of subjective quality of experience are a challenging and still unexplored area of research. We plan to continue studying the various aspects of human responses when viewing videos streamed under realistic network conditions since better models of these responses could greatly benefit future efforts to improve network streaming and encoding strategies adopted by content providers. Objective prediction models that incorporate spatio-temporal aspects of videos and that predict human reactions to both

bitrate dynamics and re-buffering events could ultimately help video engineers address resource allocation problems more efficiently and in a user-adaptive way. Recent efforts [41], [54] are important early steps towards this research goal. In the future, we plan to extend these works by focusing on continuous time QoE monitoring.

Bibliography

[1] "Cisco visual networking index: Global mobile data traffic fore- cast update, 2010–2015, Cisco Corp., 2011."

[2] "https://bitmovin.com/mpeg-dash/."

[3] T.-Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson, "A buffer-based approach to rate adaptation: Evidence from a large video streaming service," *ACM SIGCOMM Computer Communication Review*, vol. 44, no. 4, pp. 187–198, 2015.

[4] D. S. Hands and S. Avons, "Recency and duration neglect in subjective assessment of television picture quality," *Applied Cognitive Psychology*, vol. 15, no. 6, pp. 639–657, 2001.

[5] A. K. Moorthy and A. C. Bovik, "Visual quality assessment algorithms: What does the future hold?" *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 675–696, 2011.

[6] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, June 2010.

[7] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.

[8] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," in *IEEE International Conference on Image Processing*, 2011, pp. 2505–2508.

[9] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana, "Video quality assessment on mobile devices: Subjective, behavioral and objective studies," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 6, pp. 652–671, 2012.

[10] J. Y. Lin, R. Song, C.-H. Wu, T. Liu, H. Wang, and C.-C. J. Kuo, "Mcl-v: A streaming video quality assessment database," *Journal of Visual Communication and Image Representation*, vol. 30, pp. 1–9, 2015.

[11] Z. Duanmu, A. Rehman, K. Zeng, and Z. Wang, "Quality-of-experience prediction for streaming video," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[12] N. Staelens, J. D. Meulenaere, M. Claeys, G. V. Wallendael, W. V. den Broeck, J. D. Cock, R. V. de Walle, P. Demeester, and F. D. Turck, "Subjective quality assessment of longer duration video sequences delivered over http adaptive streaming to tablet devices," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 707–714, Dec 2014.

[13] D. Ghadiyaram, A. C. Bovik, H. Yeganeh, R. Kordasiewicz, and M. Gallant, "Study of the effects of stalling events on the quality of experience of mobile streaming videos," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 989–993.

[14] M. Seufert, M. Slanina, S. Egger, and M. Kottkamp, "To pool or not to pool': A comparison of temporal pooling methods for http adaptive video

streaming," in *Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, 2013, pp. 52–57.

[15] N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli, J. Astola, and V. Lukin, "On between-coefficient contrast masking of DCT basis functions," in *International Workshop on Video Processing and Quality Metrics*, vol. 4, 2007.

[16] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Proc.*, vol. 13, no. 4, pp. 600–612, 2004.

[17] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems and Computers*, 2003, pp. 1398–1402.

[18] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: a highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2014.

[19] M. H. Pinson, L. K. Choi, and A. C. Bovik, "Temporal video quality model accounting for variable frame delay distortions," *IEEE Transactions on Broadcasting*, vol. 60, no. 4, pp. 637–649, 2014.

[20] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (fvqa) index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2014, pp. 1–5.

[21] K. Manasa and S. S. Channappayya, "An optical flow-based full reference video quality assessment algorithm," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2480–2492, 2016.

[22] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 684–694, 2013.

[23] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[24] ——, "A visual information fidelity approach to video quality assessment," in *The First International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2005, pp. 23–25.

[25] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *IEEE Int'l Conf. Acoust., Speech and Signal Process., Honolulu, HI*, June 2007.

[26] A. C. Bovik, "Automatic prediction of perceptual image and video quality," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 2008–2024, 2013.

[27] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, 2013.

[28] K.-C. Yang, C. C. Guest, K. El-Maleh, and P. K. Das, "Perceptual temporal quality metric for compressed video," *IEEE Transactions on Multimedia*, vol. 9, no. 7, pp. 1528–1535, 2007.

[29] F. Yang, S. Wan, Y. Chang, and H. R. Wu, "A novel objective no-reference metric for digital video quality assessment," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 685–688, 2005.

[30] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *IEEE International Conference on Image Processing*, 2008, pp. 385–388.

[31] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing*, vol. 23, no. 3, pp. 1352–1365, 2014.

[32] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 289–300, 2016.

[33] M. Fiedler, T. Hossfeld, and P. Tran-Gia, "A generic quantitative relationship between quality of experience and quality of service," *IEEE Network*, vol. 24, no. 2, pp. 36–41, 2010.

[34] H. Yeganeh, R. Kordasiewicz, M. Gallant, D. Ghadiyaram, and A. Bovik, "Delivery quality score model for internet video," in *Image Processing (ICIP), 2014 IEEE International Conference on*, Oct 2014, pp. 2007–2011.

[35] D. Ghadiyaram, J. Pan, and A. C. Bovik, "A time-varying subjective quality model for mobile streaming videos with stalling events," in *SPIE Optical Engineering+ Applications*, 2015, pp. 959 911–959 911–8.

[36] A. Balachandran, V. Sekar, A. Akella, S. Seshan, I. Stoica, and H. Zhang, "Developing a predictive model of quality of experience for internet video," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4. ACM, 2013, pp. 339–350.

[37] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz, "Quantification of YouTube QoE via crowdsourcing," in *IEEE International Symposium on Multimedia*, 2011, pp. 494–499.

[38] D. Z. Rodriguez, J. Abrahao, D. C. Begazo, R. L. Rosa, and G. Bressan, "Quality metric to assess video streaming service over tcp considering temporal location of pauses," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 985–992, 2012.

[39] R. K. Mok, E. W. Chan, and R. K. Chang, "Measuring the quality of experience of http video streaming," in *12th IFIP/IEEE International Symposium on Integrated Network Management (IM 2011) and Workshops*. IEEE, 2011, pp. 485–492.

[40] A. J. Greene, C. Prepscius, and W. B. Levy, "Primacy versus recency in a quantitative model: activity is the critical distinction," *Learning and Memory*, vol. 7, no. 1, pp. 48–57, 2000.

[41] Z. Duanmu, Z. Kai, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE Journal of Selected Topics in Signal Processing*, 2016, to appear.

[42] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H. 264/AVC," in *IEEE Consumer Communications and Networking Conference*, 2012, pp. 127–131.

[43] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[44] *ITU-T Recommendation P.910: Subjective video quality assessment methods for multimedia applications*, Int. Telecommuncation Union Std.

[45] *BT-500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures*, Int. Telecommuncation Union Std.

[46] C. Chen, L. K. Choi, G. de Veciana, C. Caramanis, R. W. Heath, and A. C. Bovik, "Modeling the time-varying subjective quality of HTTP video streams with rate adaptations," *IEEE Trans. on Image Proc.*, vol. 23, no. 5, pp. 2206–2221, 2014.

[47] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in *KDD workshop*, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[48] J. W. Tukey, *Exploratory Data Analysis*, 1977.

[49] R. Sakia, "The Box-Cox transformation technique: a review," *The Statistician*, pp. 169–178, 1992.

[50] M. Hubert and E. Vandervieren, "An adjusted boxplot for skewed distributions," *Computational Statistics and Data Analysis*, vol. 52, no. 12, pp. 5186–5201, 2008.

[51] J. De Cock, Z. Li, M. Manohara, and A. Aaron, "Complexity-based consistent-quality encoding in the cloud," in *IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1484–1488.

[52] C. Fenimore, J. Libert, and S. Wolf, "Perceptual effects of noise in digital video compression," *SMPTE Journal*, vol. 109, no. 3, pp. 178–187, 2000.

[53] "Daala codec. https://git.xiph.org/daala.git/."

[54] C. G. Bampis, Z. Li, and A. C. Bovik, "Learning to Predict Streaming Video QoE: Distortions Rebuffering and Memory," *in preparation.*

Vita

Christos Bampis received the M.Eng. Diploma in E.E. from the National Technical University of Athens (NTUA) in 2014. He is currently pursuing the Ph.D. degree with the Laboratory for Image and Video Engineering (LIVE) at The University of Texas at Austin. His research interests include image and video quality assessment with an emphasis on adaptive video streaming.

Contact: cbampis@gmail.com

This thesis was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.